STA-380 Exercises

Varun Kausika

2022-08-10

1. Probability practice

Part A

Given information:

Two categories of users:

1. Truthful clicker (TC)
2. Random clicker (RC)

Information on probablities:

- $P(RC) = 0.3$
- $P(Yes|RC) = 0.5$
- $P(No|RC) = 0.5$
- $P(TC) = 0.7$
- $P(Yes|TC) = x$
- $P(No|TC) = 1 - x$
- $P(Yes) = 0.65$
- $P(No) = 0.35$

Using the Rule of Total Probability,

$$P(Yes) = P(Yes, TC) + P(Yes, RC) = P(TC) * P(Yes|TC) + P(RC) * P(Yes|RC) \tag{1}$$

$$P(Yes) = 0.7x + 0.3 * 0.5 = 0.7x + 0.15 = 0.65$$

Solving for x, we get,

$x = P(Yes|TC) = 0.714$

Part B

We are being asked $P(Diseased|Positive)$

Given information:

- $P(Positive|Diseased) = 0.993$
- $P(Negative|NotDiseased) = 0.9999$
- $P(Diseased) = 0.000025$

According to Bayes Rule and Rule of Total Probability,

$$P(Diseased|Positive) = \frac{P(Positive|Diseased) * P(Diseased)}{P(Positive)} \tag{2}$$

and,

$$P(Positive) = P(Positive|Diseased) * P(Diseased) + P(Positive|Not\ Diseased) * P(Not\ Diseased) \tag{3}$$

Therefore,

$$P(Positive) = 0.993 * 0.000025 + 0.0001 * 0.999975 = 0.000125$$

Substituting in (2) we get,

$$P(Diseased|Positive) = \frac{0.993 * 0.000025}{0.000125} = 0.1986$$

2. Wrangling the Billboard Top 100

Part A

First, we load in the data and perform a group by on the performer and the song, with an agg function of count for the week

Table 1: Billboards

performer	song	count
'N Sync	(God Must Have Spent) A Little More Time On You	22
'N Sync	Bye Bye Bye	23
'N Sync	Gone	24
'N Sync	I Drive Myself Crazy	12
'N Sync	I Want You Back	24
'N Sync	It's Gonna Be Me	25
'N Sync	Pop	15
'N Sync	Tearin' Up My Heart	1
'N Sync	This I Promise You	26
'N Sync & Gloria Estefan	Music Of My Heart	20

Finally, we sort the dataframe in descending order of counts and find the top 10 and give our table a caption:

Table 2: Top 10 most popular songs

performer	song	count
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
Jason Mraz	I'm Yours	76
The Weeknd	Blinding Lights	76
LeAnn Rimes	How Do I Live	69

performer	song	count
LMFAO Featuring Lauren Bennett & GoonRock	Party Rock Anthem	68
OneRepublic	Counting Stars	68
Adele	Rolling In The Deep	65
Jewel	Foolish Games/You Were Meant For Me	65
Carrie Underwood	Before He Cheats	64

Part B

First we group by year and order by ascending year. Then, we remove the years 1958 and 2021 from the rows and order just to make sure. Finally, we proceed to plot the columns.

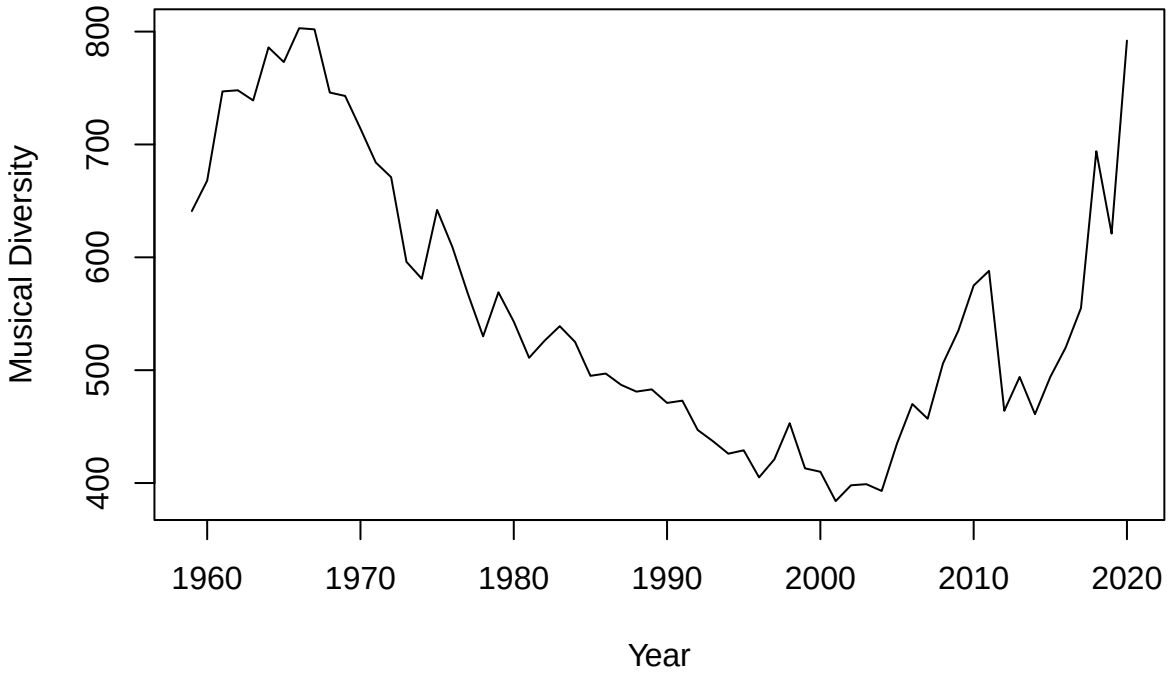

Figure 1: The plot shows peaks in diversity in 1965 and 2020, along with extreme lows in 2000

Part C

First, we filter the dataframe from part A to include only those songs with weeks at least 10. Then, We do a group by on the artists. Finally, we can select those artists with a hit-count of at least 30.

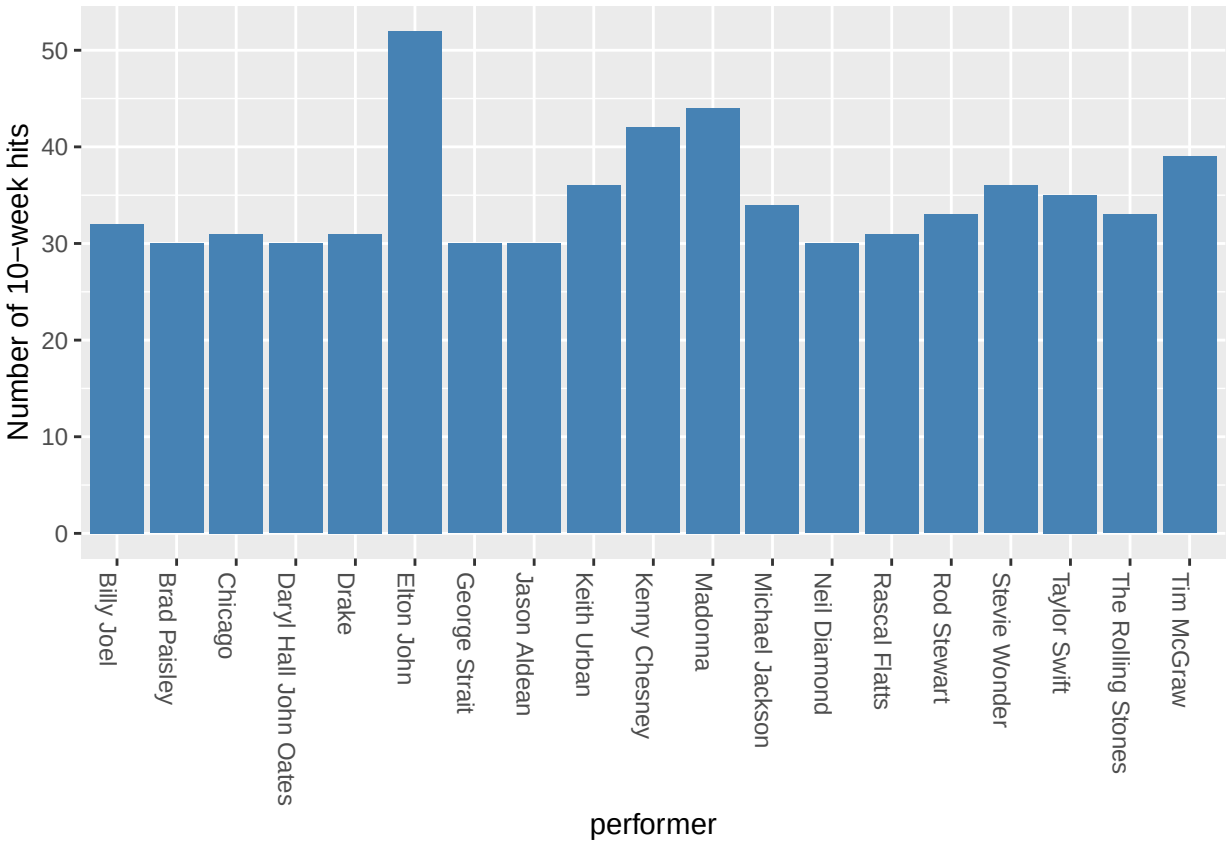

Figure 2: Elton John has more hits than others by quite a large margin